IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received September 18, 2021, accepted October 10, 2021, date of publication October 15, 2021, date of current version October 26, 2021.

Digital Object Identifier 10.1109/ACCESS.2021.3120710

Semi-Supervised Anomaly Detection Algorithm Using Probabilistic Labeling (SAD-PL)

KIBAE LEE[1], **(Graduate Student Member, IEEE),**
CHONG HYUN LEE[1], **(Member, IEEE), AND JONGKIL LEE**[2]
[1]Department of Ocean System Engineering, Jeju National University, Jeju 63243, South Korea
[2]Department of Mechanical Engineering Education, Andong National University, Andong 36729, South Korea

Corresponding author: Chong Hyun Lee (chonglee@jejunu.ac.kr)

This work was supported in part by the Defense Acquisition Program Administration, and in part by the Agency for Defense Development in Korea under Contract UD200011DD.

ABSTRACT To detect abnormal data via semi-supervised learning, unlabeled data are generally assumed to be normal data. This assumption, however, causes inevitable performance degradation when a small fraction of abnormal data is included in the unlabeled dataset. To overcome the degradation and to maintain stable detection performance, we propose a semi-supervised anomaly detection algorithm using probabilistic labeling (SAD-PL) for unlabeled data. The proposed SAD-PL is composed of two steps: (1) estimating local outlier factor (LOF) scores of latent vectors from both labeled and unlabeled data and (2) estimating labeling probability on the unlabeled data by using the prior missing probability of the labeled data via the Neyman-Pearson (NP) criterion. The SAD-PL runs iteratively by using the proposed complementary learning functions until the rate of label changes is lower than the predefined threshold. Experimental results reveal that the SAD-PL shows superior detection probability over the existing algorithms and stable performance regardless of the normal to abnormal data ratio in unlabeled data and the ratio of change variation of unlabeled data statistics to labeled data statistics.

INDEX TERMS Anomaly detection, semi-supervised learning, probabilistic labeling, Neyman-Pearson criterion, local outlier factor.

I. INTRODUCTION

Anomaly detection is used for detecting abnormal samples that deviate from the predefined normality [1]. It has various applications, such as in medicine, security, and manufacturing [1]. Further applications include intrusion detection in cybersecurity [2]–[4], industrial fault and damage detection in monitoring sensor data [5]–[7], and acoustic novelty detection for audio surveillance and underwater sonar systems [8]–[13]. Typical anomaly detection methods based on unsupervised learning assume that most of the samples are normal. The unsupervised approaches, primarily treated as a one class classification problem, learn features of normal samples [14]–[18]. Typical anomaly detection methods such as the one class support vector machine (OC-SVM) [14] and support vector data description (SVDD) [15] attempt to learn compact descriptions of normal samples. Recent deep learning approaches have shown outstanding performance by overcoming the problems with shallow learning on high-dimensional data [16]–[18]. Deep SVDD [16], a representative deep approach, trains a neural network, while minimizing the volume of a hypersphere that encloses normal samples in latent space. Recent algorithms including Deep Multi-sphere SVDD [17] and deep robust one-class classification (DROCC) [18], have been proposed to learn representation of normal samples. Most unsupervised approaches that are not trained on abnormalities have limited detection performance.

Some labeled data, as well as unlabeled data, may be utilized in real-world applications, and an especially small number of anomalous samples in labeled data can be used. Song *et al.* [19] and Akvay *et al.* [20] proposed semi-supervised anomaly detection models that use reliable normal samples in unlabeled data for training. However, since these models do not train the abnormalities like the unsupervised approaches, they have limited performance. Ruff *et al.* [21] proposed a deep semi-supervised anomaly detection (Deep SAD) that learns anomalous samples in labeled data. By assuming that most unlabeled samples are normal, the Deep SAD trains the normal data to be concentrated in the center of the latent space, and then trains

The associate editor coordinating the review of this manuscript and approving it for publication was Mehul S. Raval.

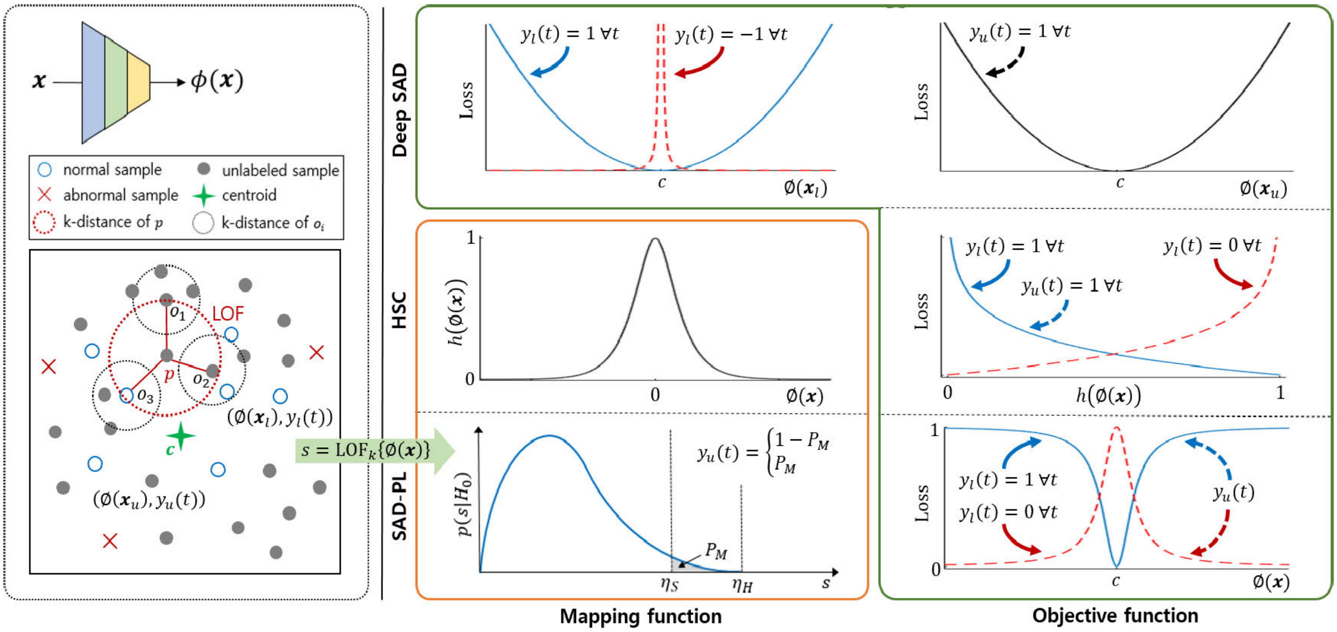


FIGURE 1. Overview of the different approaches to semi-supervised anomaly detection. Left: Semi-supervised anomaly detection models aim to separate normal and abnormal samples in latent space. Deep SAD learns normal and abnormal samples with asymmetric objectives. HSC is trained by complementary losses based on radial basis function. Deep SAD and HSC is trained assuming that the samples in unlabeled data are normal. SAD-PL learns both labeled and unlabeled data with complementary objectives depending on labeling probability.

labeled abnormal samples to move away from the center of the latent space. An unsupervised Outlier Exposure (OE) approach for learning OE data and unlabeled data has been proposed based on a similar assumption of semi-supervised anomaly detection [22]. The anomaly detection method based on unsupervised OE learning trains a binary classifier on the OE data and the unlabeled abnormal and normal data. Hyper sphere classification (HSC) [23] is a class classification algorithm based on unsupervised OE learning that uses the relative distance in latent space for training. The Deep SAD and HSC are advanced anomaly detection algorithms since they consider abnormal samples or OE data in the training process. However, their assumption of most unlabeled samples being normal inevitably causes performance degradation as the number of abnormal samples in the unlabeled dataset increases. To overcome the degradation, a new semi-supervised anomaly detection algorithm that can learn unlabeled normal and abnormal data efficiently is required.

In this paper, we propose an anomaly detection algorithm based on probabilistic normal or abnormal labeling for each sample in unlabeled data. The proposed algorithm, denoted as the semi-supervised anomaly detection algorithm using probabilistic labeling (SAD-PL), involves the two-step probabilistic labeling process: (1) computing the local outlier factor (LOF) score of latent vectors from both labeled and unlabeled data and (2) estimating the labeling probability on the unlabeled data by using the missing probability of labeled normal data via the Neyman-Pearson (NP) criterion. The SAD-PL runs until the labeling change rate becomes lower than the preset threshold.

Our paper is organized follows: Section 2 describes the SAD-PL and evaluates the performance using toy data. In Section 3, the experimental results on the image dataset are described in comparison with existing algorithms. Conclusions are presented in Section 4.

II. SAD-PL

In this section, we introduce the SAD-PL based on semi-supervised learning with probabilistic labeling. Figure 1 shows the proposed algorithm along with the existing semi-supervised algorithms of the Deep SAD and HSC. The proposed SAD-PL uses feature representations $\phi(\boldsymbol{x})$ through autoencoder pretraining and learns the encoded normal samples to close centroid c in latent space without the decoding network. Then, the SAD-PL is trained according to the proposed probabilistic labeling, which uses the LOF score $s = \mathrm{LOF}_k\{\phi(\boldsymbol{x})\}$. The LOF scoring the density based on the relative distance between k neighbor samples is known to show robust performance in the multimodal normality case [24]. As shown in Figure 1, the LOF score of the object p is computed as follows: (1) computing k-distances of p with the k neighbor samples o_i (2) computing k-distances of o_i with the k neighbor samples (3) computing ratio of average of the k-distances obtained by (2) to the k-distances by (1). To detect a small number of group anomaly samples, the SAD-PL sets k large enough to cover the relative distance between normal and abnormal samples. The NP criterion is used to determine the threshold that satisfies the detection or missing probabilities under a given constraint [25]. The probability used for labeling is computed by using the missing probability of the labeled normal data P_M as

$$P_M = \int_{\eta}^{\infty} p\left(s \mid H_0\right) ds \tag{1}$$

where $p\left(s \mid H_0\right)$ denotes the probability density function of $s = \text{LOF}_k\{\phi\left(\boldsymbol{x}\right)\}$ obtained from labeled normal data (H_0). η denotes the threshold for the given P_M. Note that η varies as the SAD-PL. These changes in η cause the probabilistic label $y(t)$ to change each training epoch t. However, labels $y_l(t)$, $l \leq n$ for n labeled samples $\boldsymbol{x}_l$ are fixed as follows:

$$y_l\left(t\right) = \begin{cases} 1, & \forall t \text{ for normal} \\ 0, & \forall t \text{ for abnormal} \end{cases} \tag{2}$$

For unlabeled data $\boldsymbol{x}_u$, the labels $y_u\left(t\right), u > n$ are defined as follows:

$$y_u(t) = \begin{cases} 1 - P_M, & s_u \leq \eta \\ P_M, & s_u > \eta \end{cases} \tag{3}$$

where s_u represents the LOF score of $\boldsymbol{x}_u$. According to P_M determined by the NP criterion, $y_u\left(t\right)$ has a corresponding probability and consequently implies the probabilistic label for the unlabeled $\boldsymbol{x}_u$. For network ϕ training, the SAD-PL uses datasets $\{\boldsymbol{x}_l, y_l\left(t\right), l \leq n\}$ and $\{\boldsymbol{x}_u, y_u\left(t\right), u > n\}$, which are composed of n labeled and $(m - n)$ unlabeled data from the corresponding probabilistic labels estimated in (2) and (3). The resultant objective function of the SAD-PL can be described as follows:

$$\begin{aligned} \min_{w} \; & \frac{1}{m} \sum_{i=1}^{m} y_i(t) d\left(\phi\left(\boldsymbol{x}_i\right)\right) \\ & + \left(1 - y_i(t)\right)\left(1 - d\left(\phi\left(\boldsymbol{x}_i\right)\right)\right) + \frac{\lambda}{2} \sum_{j=1}^{J} \left\| \boldsymbol{W}^j \right\|_F^2 \end{aligned} \tag{4}$$

where $\boldsymbol{W}^j$ is the weights of layer $j \in \{1, \cdots, J\}$, $\|\cdot\|_F$ denotes the Frobenius norm and $d(\phi\left(\boldsymbol{x}_i\right))$ is the function of the distance from $\boldsymbol{c}$ using Geman-McClure loss and defined as follows:

$$d\left(\phi\left(\boldsymbol{x}_i\right)\right) = \frac{\|\phi\left(\boldsymbol{x}_i\right) - \boldsymbol{c}\|^2}{\|\phi\left(\boldsymbol{x}_i\right) - \boldsymbol{c}\|^2 + 1} \tag{5}$$

The proposed SAD-PL objective function consists of a weight decay regularizer on $\boldsymbol{W}^j$ multiplied by hyperparameter $\lambda > 0$ for preventing overfitting and two complementary learning functions of symmetrical losses for normal and abnormal samples, which are multiplied by complementary probabilistic labels $y_i(t)$. The SAD-PL learns the normal samples closer to $\boldsymbol{c}$ and the abnormal samples away from $\boldsymbol{c}$ for the labeled data. For unlabeled data training, the network uses complementary losses with probabilistic labeling $y_u(t)$. To avoid a trivial solution of $\mathcal{W} = 0$, the bias is not updated during the learning process. By adopting the Geman-McClure loss in (5), the SAD-PL can prevent divergence in the learning process with a limited loss of 0 to 1 and obtain robust stability for mislabeled data [26]. Note that the SAD-PL object function has both regression and classification properties since it learns by computing distance loss based on the estimated probabilistic labels $y_u\left(t\right)$ for unlabeled data.

Soft Labeling: To find the optimum P_M for unlabeled data labeling, the SAD-PL adopts the ensemble networks $[\phi_1, \cdots, \phi_Q]$, where ϕ_q is the model trained with $P_M(q)$. To estimate the optimum P_M, the SAD-PL uses the fact that the LOF difference between normal and abnormal samples in well-trained networks becomes relatively larger than that in ill-trained networks. To determine the optimum $P_M(q)$, we compute the LOF difference $\Delta s\left(q\right)$ for ϕ_q as follows:

$$\Delta s(q) = \left|\bar{s}\left(q\right)_0 - \bar{s}\left(q\right)_1\right| \tag{6}$$

where $\bar{s}\left(q\right)_0$ and $\bar{s}\left(q\right)_1$ represent the average LOFs of the labeled normal and abnormal samples, respectively. Then, from Q differences $[\Delta s\left(1\right), \cdots, \Delta s\left(Q\right)]$, the optimum $P_M(o)$ is determined by finding the maximum $\Delta s(q)$ at a given epoch T. The SAD-PL learns the model by using the threshold η_S obtained from $P_M(o)$ and the estimated probabilistic labeling procedure described above.

Hard Labeling: By setting $P_M(o) = 0$, the SAD-PL can reduce the computational cost of memory and time. This version of the SAD-PL is known as hard labeling and learns the network by using the threshold denoted as η_H in Figure 1.

The proposed SAD-PL runs until the labeling change rate $\Delta y_u(t)$ at iteration t and is lower than the preset threshold ε. The $\Delta y_u(t)$ is computed as

$$\Delta y_u\left(t\right) = \frac{1}{m - n} \sum_{u=n+1}^{m} \left| \frac{y_u\left(t\right) - y_u(t-1)}{1 - 2P_M(o)} \right| \tag{7}$$

Algorithm 1 shows the overall learning procedure of the SAD-PL.

The proposed SAD-PL is related to the Deep SAD algorithm. Setting $y_l(t) = 1$ for the normal samples, $y_l(t) = -1$ for the abnormal samples, $y_u\left(t\right) = 1 \forall t$, setting $d\left(\phi\left(\boldsymbol{x}_i\right)\right) = \|\phi\left(\boldsymbol{x}_i\right) - \boldsymbol{c}\|^2$ and $d\left(\phi\left(\boldsymbol{x}_i\right)\right)$ to $y_l(t)$ power in Equation (4), gives the Deep SAD object function as follows [21]:

$$\begin{aligned} \min_{w} \; & \frac{\zeta}{m} \sum_{l=1}^{n} \left(\|\phi\left(\boldsymbol{x}_l\right) - \boldsymbol{c}\|^2\right)^{y_l(t)} \\ & + \frac{1}{m} \sum_{u=n+1}^{m} \|\phi\left(\boldsymbol{x}_u\right) - \boldsymbol{c}\|^2 + \frac{\lambda}{2} \sum_{j=1}^{J} \left\| \boldsymbol{W}^j \right\|_F^2 \end{aligned} \tag{8}$$

where the hyperparameter ζ controls the balance of learning between labeled and unlabeled terms.

Additionally, the proposed SAD-PL is related to the HSC model with the objective function of cross entropy for relative distance in latent space [23]. By taking the negative and logarithmic distance losses in Equation (4) and replacing $d(\phi\left(\boldsymbol{x}_i\right))$ with the radial basis function $h(\phi\left(\boldsymbol{x}_i\right)) = \exp\{-(\sqrt{\|\phi\left(\boldsymbol{x}\right)\|^2 + 1} - 1)\}$ gives the HSC object function

Algorithm 1 Learning Procedure of SAD-PL

```
Input:
  Labeled data: $(\boldsymbol{x}_l, y_l)$
  Unlabeled data: $\boldsymbol{x}_u$
  Number of neighbors: $k$
  Increment size of $P_M$: $\delta$
  Threshold for labeling change rate: $\varepsilon$
  Mode: “hard labeling” or “soft labeling”
Output:
  Trained model: $\phi$
  Probabilistic labels: $y_u$
Initialize:
  Pretrain autoencoder: $\phi_0$
  Compute centroid: $\boldsymbol{c}$
  Create $Q$ models: $[\phi_1, \cdots, \phi_Q]$
if mode is “hard labeling” then $P_M(o) = 0$
else if mode is “soft labeling”
  for $t = 1, 2, \cdots, T$ do
    for $q = 1, 2, \cdots, Q$ do
      if $q = 1$ then $P_M(q) = 0$
      else $P_M(q) = P_M(q-1) + \delta$
      Compute LOF score: $s(q) = \mathrm{LOF}_k(\phi_q(\boldsymbol{x}))$
      Compute $\Delta s(q)$
      Threshold estimation: $\eta(q)$
      Probabilistic labeling: $y_u(t, q)$
      Train model: $\phi_q$
    end
  end
  if $\Delta s(o)$ is maximum then
    Determine $P_M = P_M(o)$
    Set $\phi = \phi_o$
  end
end
for $t = 1, 2, \cdots$ do
  Compute LOF score: $s = \mathrm{LOF}_k(\phi(\boldsymbol{x}))$
  Threshold estimation: $\eta$
  Probabilistic labeling: $y_u(t)$
  Train model: $\phi$
  if $\Delta y_u(t) < \varepsilon$ then break
end
```

as follows [23]:

$$\min_{w} -\frac{1}{m}\sum_{i=1}^{m} y_i(t)\log h(\phi(\boldsymbol{x}_i)) + (1-y_i(t))\log\{1-h(\phi(\boldsymbol{x}_i))\} + \frac{\lambda}{2}\sum_{j=1}^{J}\left\|\boldsymbol{W}^j\right\|_F^2 \quad (9)$$

where $y_i(t) = 1 \forall t, i \in \{n+1, \cdots, m\}$ for unlabeled data.

A. COMPARISON OF ANOMALY DETECTION MODELS

We compare the anomaly detection models described above with toy examples. The training data consist of a total of 10000 samples, of which 10% are labeled data and 90% are unlabeled data. Five percent of labeled data and 1% of unlabeled data consist of abnormal samples. We generate the normal samples in the training data as two-dimensional big moon and small moon patterns. We also add Gaussian noise of variance 0.2 and 0.25 to the labeled and unlabeled data, respectively. The abnormal samples in labeled data are located to be clearly distinguished from the normal samples. On the other hand, the abnormal samples in unlabeled data are located adjacent to the boundaries of the normal samples. The test data consist of 1000 normal and abnormal samples. The normal samples contain Gaussian noise with a variance of 0.3 in big moon and small moon patterns. The abnormal samples are generated with a uniform distribution.

The models for comparison use the same encoding network of the pretrained autoencoder. The encoding network consists of two hidden layers of 100 nodes, followed by ELU activations that represent a two-dimensional input in a two-dimensional latent space. For autoencoder pretraining, we employ the above architectures for the encoding networks and then construct decoding networks symmetrically. We set $\zeta = 1$ for the Deep SAD, and $k = 100$ and $\varepsilon = 0.0001$ for the SAD-PL. We use $P_M = 0.04$ obtained from soft labeling training. For all models, we set $\lambda = 10^{-6}$ and use the Adam optimizer with a learning rate of 10^{-5}. In addition, we train all models except the SAD-PL in epoch 300. Figure 2 shows the decision boundaries with training data and the test AUC (area under the receiver operating characteristic curve) of anomaly detection models. The decision boundaries in Figure 2 are represented with an upper bound on 10% of the anomaly score normalized via min-max scaling. The anomaly score is computed $\|\phi(\boldsymbol{x}) - \boldsymbol{c}\|^2$ for the Deep SAD and the SAD-PL and $\|\phi(\boldsymbol{x})\|^2$ for HSC. In (a) and (c) of Figure 2, the Deep SAD and HSC create the decision boundaries along the perimeter of the large moon and small moon patterns because the unlabeled data are assumed to be normal. Therefore, the abnormal samples close to the normal samples in unlabeled data are located within the decision boundaries, and the decision boundaries are created wider in a region where the abnormal sample in the labeled data does not exist. However, the SAD-PL represents the tight decision boundaries with complementary learning for the labeling probability of the unlabeled data in (e) of Figure 2. In particular, the SAD-PL has higher anomaly scores than the existing methods in areas separated inside the large moon and small moon patterns. In (b) of Figure 2, the supervised deep SAD is trained with only normal samples in unlabeled data, since the labels are not used on the unlabeled term in the objective function. For this reason, the supervised Deep SAD represents compact decision boundaries compared to the semi-supervised Deep SAD. However, the supervised Deep SAD, which learns relatively few abnormal samples, creates wider decision boundaries than the SAD-PL. (d) and (f) in Figure 2 show the decision boundaries and test AUC for the supervised HSC and SAD-PL learning the unlabeled data with labels. The supervised HSC and SAD-PL have extremely tight decision boundaries and high AUCs of 99.25% and 99.28%, respectively. These

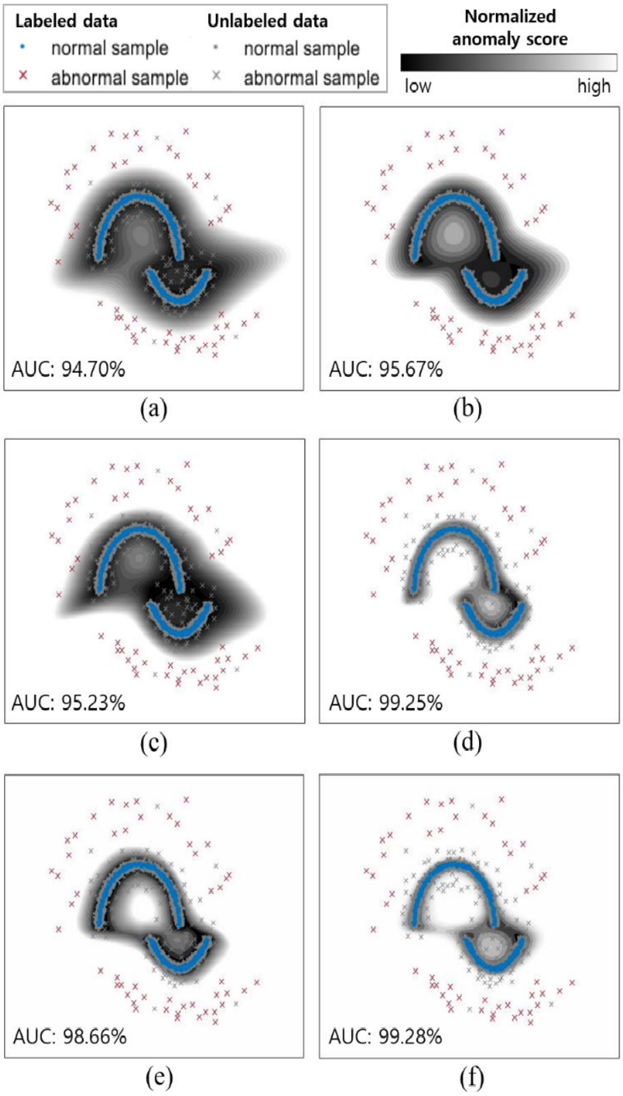


FIGURE 2. The decision boundaries and test AUC of anomaly detection models. (a) Semi-supervised Deep SAD, (b) Supervised Deep SAD, (c) Semi-supervised HSC, (d) Supervised HSC, (e) Semi-supervised SAD-PL, (f) Supervised SAD-PL.

results show improved AUCs of more than 4.02% compared to the semi-supervised HSC, but only up to a 0.62% difference in AUCs compared to the semi-supervised SAD-PL.

Figure 3 shows the training AUCs of the comparative models according to epoch and the $\Delta y_u(t)$ in the SAD-PL learning. In Figure 3, the HSC, which learns distance based on a radial basis function, achieves higher learning efficiency than the Deep SAD. However, the HSC and the Deep SAD, assuming the unlabeled data are normal, show slower learning caused by adversarial learning between the abnormal samples in labeled and unlabeled data. However, the SAD-PL achieves high learning efficiency with complementary learning based on probabilistic labeling of unlabeled data. We can also predict the completion of learning as $\Delta y_u(t)$ converges to zero as the epoch increases.

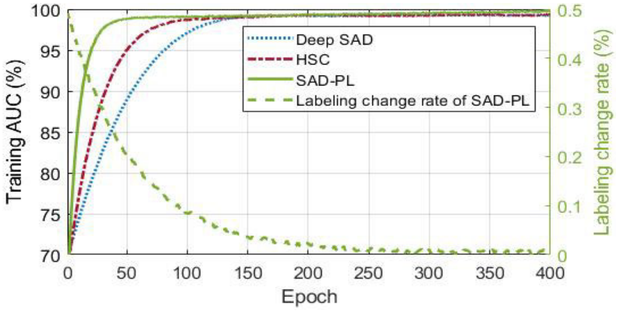


FIGURE 3. Training AUC of anomaly detection models according to epoch and change rate of probabilistic labels in SAD-PL model.

B. CHARACTERISTICS OF PROBABILISTIC LABELING

The SAD-PL sets the labeling probability on unlabeled data with P_M of the labeled data via the NP criterion. It determines η by Equation (1) and sets the labeling probability by (3). Typically, unlabeled data consist of a large number of samples with various statistics, such as variance, rather than labeled data. Therefore, the NP condition may not be met on unlabeled data for the η set in labeled data. The proposed algorithm repeats the probabilistic labeling in an iteration of learning to change normal samples in labeled and unlabeled data into a similar probability distribution of LOF scores. In this learning process, the missing probability for the normal unlabeled samples converges to the predefined P_M. Figure 4 shows histograms of LOF scores for the normal samples in an iteration of training. In (a) and (b) of Figure 4, the probability distributions of LOF scores for normal samples in labeled and unlabeled data become similar with each iteration of training. As a result, the missing probability for the normal unlabeled samples varies from 20.25% to 4.02%, close to the predefined $P_M = 0.04$. Figure 5 shows that the labeling accuracy for unlabeled data, which is mostly composed of normal samples, reaches approximately 96%.

To verify the soft probabilistic labeling, we compute the difference in LOF scores $\Delta s(q)$ according to epoch. Figure 6 shows three $\Delta s(q)$s according to three $P_M(q)s$ along with test AUCs. As shown in Figure 6, we can observe that a larger $\Delta s(q)$ corresponds to a larger AUC, and thus, can determine a P_M suitable for unlabeled data via the proposed soft probabilistic labeling procedure.

III. EXPERIMENTS

We evaluate the SAD-PL on the well-known MNIST [27], CIFAR10 [28], and MNIST-C [29] datasets. The SAD-PL is compared with the methods based on one class classification. We present results from unsupervised methods of SVDD [15] and the Deep SVDD [16] and semi-supervised methods of the Deep SAD [21] and HSC [23]. We implement the semi-supervised HSC by replacing OE data with the labeled data. We also present the results of the supervised Deep SAD, HSC, and SAD-PL by using labeling information of the unlabeled data. The supervised Deep SAD learns only normal samples in unlabeled data since the labels are not

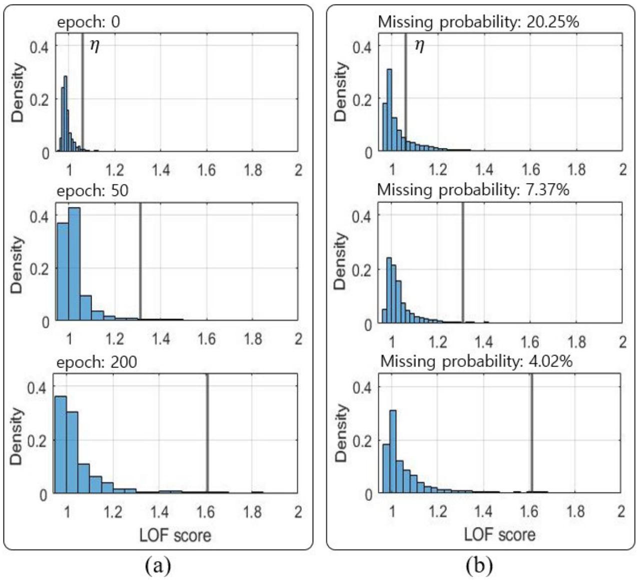


FIGURE 4. Histogram for LOF score according to epochs (a) Normal samples of labeled data (b) Normal samples of unlabeled data.

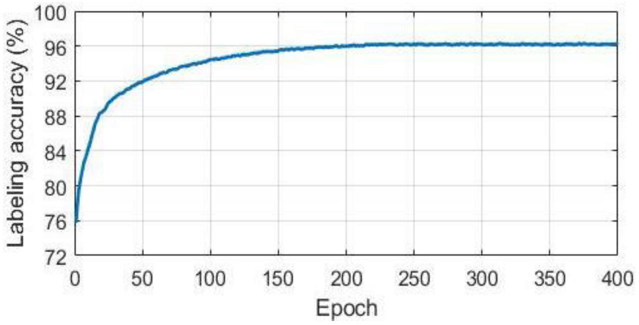


FIGURE 5. Labeling accuracy of SAD-PL training according to epoch with $P_M = 0.04$.

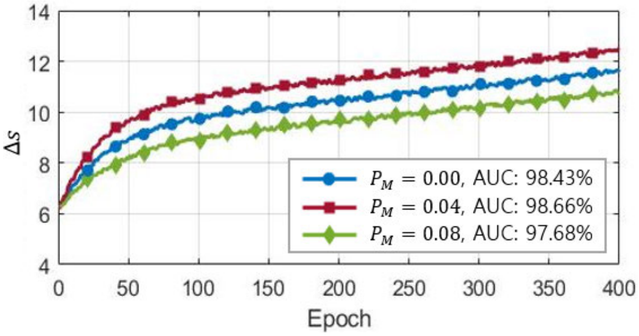


FIGURE 6. Difference of LOF score and test AUC between normal and abnormal samples in labeled data according to P_M and epoch.

used on unlabeled term in the objective function. We run all experiments for $\nu \in \{0.1, 0.25, 0.5\}$ of SVDD with a Gaussian kernel and show the corresponding results. The deep models use the same encoding network structure in the pretrained autoencoder. We employ LeNet-type convolutional neural networks (CNNs), where each convolutional module consists of a convolutional layer followed by leaky ReLU activations and 2×2 max-pooling. In the MNIST and MNIST-C experiments, we employ a CNN with two modules, $8 \times (5 \times 5)$ filters followed by $4 \times (5 \times 5)$ filters, and a final dense layer of 32 units. In the CIFAR10 experiments, we employ a CNN with three modules, $32 \times (5 \times 5)$ filters, $64 \times (5 \times 5)$ filters, and $128 \times (5 \times 5)$ filters, followed by a final dense layer of 128 units. For the pretraining autoencoder, we employ identical encoding networks and then construct the decoding networks symmetrically, where we replace max-pooling with simple upsampling and convolutions with deconvolutions. We use a batch size of 200 and set $\lambda = 10^{-6}$. We also use the Adam optimizer with a learning rate of 10^{-5}. For experiments on the Deep SVDD, we use the one-class Deep SVDD model [16]. We run all experiments for $\zeta \in \{0.01, 0.1, 1, 0, 100\}$ of the Deep SAD and show the best results. The SAD-PL is evaluated via two separate hard and soft labels. We set $k = 200$ and complete learning when $\Delta y_u(t)$ is less than $\varepsilon = 0.001$. In soft labeling, we set $Q = 101$ by setting $P_M = 0$ to $P_M = 0.2$ with an $\delta = 0.002$ interval and select the $P_M(o)$ in which Δs is maximum in $T = 10$. We train the deep models for 300 epochs. The image data used in the experiments are normalized through min-max scaling.

We use a typical one vs. rest evaluation method on the MNIST and CIFAR10 datasets [30]. On MNIST and CIFAR 10, we set the ten classes to be normal classes and let the remaining nine classes represent anomalies. We use the original training and test data. In the training data, we constitute most of the data as the normal class and replace a small amount with the data from the abnormal class according to the experimental scenario. We also divide the training data into labeled and unlabeled data, while organizing the abnormal samples in labeled data from a single anomalous class. However, the abnormal samples in unlabeled data equally contain samples of all anomalous classes. The class for abnormal samples in labeled data is randomly determined in each experiment. This gives training set sizes of approximately 6000 for MNIST and 5000 for CIFAR10. Both test sets have 10000 samples, including samples from the nine anomalous classes for each setup. On MNIST-C, we set original images to be normal and corrupted images to be abnormal. We use pre-configured training and test data. We also make the training data into most normal samples and a few abnormal samples according to the scenario. The training data are divided into labeled and unlabeled data. We organize the abnormal samples in labeled data using corrupted images of the same type. The abnormal samples in unlabeled data include all kinds of corrupted images equally. The type of corrupted images for abnormal samples in labeled data is randomly determined in each experiment. This configuration gives a training set size of approximately 60000 and a test set size of 160000 for MNIST-C.

We use 5% as labeled data and 95% as unlabeled data in the training data as Ruff *et al.* [21] do for evaluation of the Deep SAD. We constituted 98% normal samples and 2% abnormal samples in the labeled data, along with 99%

TABLE 1. Average AUCs in % for one vs. rest evaluation on MNIST dataset.

Normal class	Unsupervised learning		Semi-supervised learning				Supervised learning		
	SVDD	Deep SVDD	Deep SAD	HSC	SAD-PL (Hard)	SAD-PL (Soft)	Deep SAD	HSC	SAD-PL
0	89.91	90.38	91.84	93.59	95.22	**97.29**	92.86	99.45	99.59
1	95.39	98.30	**99.73**	99.57	98.58	99.70	99.75	99.94	99.87
2	77.08	79.95	80.51	83.06	86.72	**88.77**	84.05	89.39	98.67
3	81.64	83.10	84.34	84.49	89.64	**91.72**	85.71	98.05	99.33
4	89.09	91.49	92.26	91.73	94.34	**96.43**	93.37	98.62	99.64
5	80.96	81.38	82.31	86.72	91.09	**93.18**	84.41	99.07	99.57
6	91.50	92.77	93.82	96.07	96.48	**98.53**	94.98	98.36	99.63
7	88.36	91.10	92.55	97.28	96.61	**98.62**	94.62	98.90	99.55
8	81.74	84.12	85.04	77.23	92.45	**94.50**	86.78	98.68	99.35
9	87.36	90.24	91.73	88.59	95.32	**97.42**	93.11	98.85	99.59
Average	**86.30**	**88.28**	**89.41**	**89.83**	**93.65**	**95.62**	**90.96**	**98.83**	**99.48**

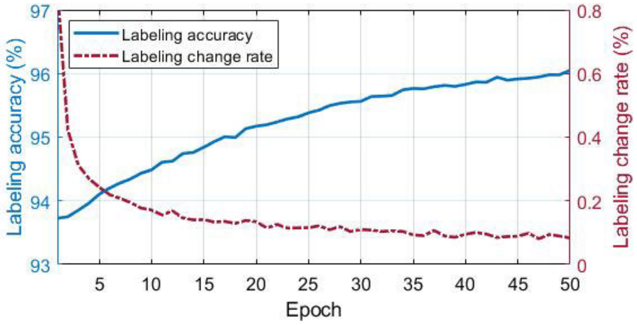


FIGURE 7. Labeling accuracy and change rate in SAD-PL training with $P_M = 0.04$ for one vs. rest evaluation(normal class: 0) on MNIST dataset.

normal samples and 1% abnormal samples in the unlabeled data. We present the evaluation results for the models with an average AUC of 30 times. Table 1 shows the evaluation results on MNIST dataset. In Table 1, the SAD-PL (Hard) and the SAD-PL (Soft) indicate the evaluation results for hard and soft labeling, respectively. The SAD-PL achieves a better performance than existing unsupervised and semi-supervised anomaly detection models with an average AUC of at least 93.65%. The SAD-PL also represents an average AUC that is at least 2.69% higher than the supervised deep SAD by training abnormal samples in unlabeled data. The semi-supervised SAD-PL has an improved average AUC of 1.97% via soft labeling, which is only 3.86% lower than that of the supervised SAD-PL. Additionally, in the experiment in which digit ‘0’ is a normal class on MNIST, the SAD-PL is trained with $P_M = 0.04$ determined by soft labeling. Figure 7 shows the labeling accuracy and $\Delta y_u(t)$ in the iteration of the SAD-PL training with soft labeling. As shown in Figure 7, the labeling accuracy reaches approximately 96%, corresponding to $P_M = 0.04$, and $\Delta y_u(t)$ converges to 0.1% or less as the epoch increases.

Tables 2 and 3 show the evaluation results for anomaly detection models on CIFAR10 and MNIST-C, respectively. In Table 2, the SAD-PL proposed achieves a better detection performance than existing unsupervised and semi-supervised methods with an average AUC of at least 64.37%. The SAD-PL with hard labeling also represents an average AUC that is 1.69% higher than the supervised deep SAD. We can see that the SAD-PL with soft labeling has a 5.22% improvement in the average AUC compared to the SAD-PL with hard labeling. This result represents an average AUC that is only 2.43% lower than the supervised SAD-PL. In Table 3, the SAD-PL similarly achieves a higher performance than the existing unsupervised and semi-supervised models, with an average AUC of at least 90.05% in the experiment using the MNIST-C dataset. The SAD-PL also represents an average AUC that is at least 5.22% higher than the supervised deep SAD. The proposed method with soft labeling has a 2.38% improvement in the average AUC compared to the SAD-PL with hard labeling.

By comparing Table 1 and 2, we can see that the SAD-PL (Soft) shows higher improvement in average AUC to the SAD-PL (Hard) when the unlabeled data set is rather complicated as CIFAR10 images. The reason of improvement is that the SAD-PL (Soft) learns the statistics of unlabeled normal data by employing the ensemble networks at the expense of longer computation time. However, if statistics of unlabeled data are similar with the labeled normal data, then SAD-PL (Hard) can show compatible performance to the SAD-PL (Soft).

Next, we investigate the effect of including labeled data during training on the MNIST, CIFAR10 and MNIST-C datasets by increasing the ratio of labeled training data from 0% to 10% and presenting the averaged AUCs of 30 times. We compute the average AUCs according to the one vs. rest evaluation method on the MNIST and CIFAR10 datasets. The ratio of abnormal samples in labeled and unlabeled data is maintained at 5% and 1%, as in the previous experiments. Figure 8 shows variations of performance in the average AUCs for the semi-supervised models according to the ratio of labeled data during training. The HSC, a classification

TABLE 2. Average AUCs in % for one vs. rest evaluation on CIFAR10 dataset.

Normal class	Unsupervised learning		Semi-supervised learning				Supervised learning		
	SVDD	Deep SVDD	Deep SAD	HSC	SAD-PL (Hard)	SAD-PL (Soft)	Deep SAD	HSC	SAD-PL
Airplane	59.82	62.56	63.64	67.98	73.70	**78.92**	64.62	80.94	79.45
Automobile	48.48	50.61	52.09	42.81	52.21	**57.29**	52.98	70.11	62.59
Bird	61.28	62.63	**64.41**	64.39	58.41	63.82	66.48	68.56	65.21
Cat	50.39	52.76	54.31	59.48	59.31	**64.52**	56.29	69.59	68.28
Deer	64.73	65.96	68.37	72.89	77.35	**82.64**	69.57	78.59	83.73
Dog	58.57	60.59	61.67	58.26	58.29	**63.47**	63.37	70.11	64.69
Frog	67.04	69.20	70.32	68.31	71.68	**76.89**	71.84	82.43	79.10
Horse	54.63	56.72	58.18	55.48	60.72	**65.98**	59.43	74.60	68.98
Ship	64.33	66.89	68.29	**76.78**	69.35	74.52	69.13	84.43	77.42
Truck	53.79	56.28	57.76	60.40	62.69	**67.87**	58.69	70.75	70.82
Average	**58.30**	**60.42**	**61.90**	**62.68**	**64.37**	**69.59**	**63.24**	**75.01**	**72.02**

TABLE 3. Average AUCs in % for evaluation on MNIST-C dataset.

Model		AUC
Unsupervised learning	SVDD	67.59
	Deep SVDD	82.79
Semi-supervised learning	Deep SAD	83.98
	HSC	85.10
	SAD-PL (Hard)	90.05
	SAD-PL (Soft)	**92.43**
Supervised learning	Deep SAD	85.04
	HSC	96.27
	SAD-PL	98.52

model, shows high improvement in the average AUC as the labeled data increase. In particular, HSC presents a higher average AUC than the Deep SAD when the ratio of the labeled data is 5% or more. The Deep SAD, a regression model that is trained with distance loss, shows robust performance, even with a small amount of labeled data. However, the SAD-PL, having both properties of regression and classification by learning with complementary objectives, represents a high average AUC compared to the existing HSC and the Deep SAD. The SAD-PL with hard labeling has a similar average AUC to the HSC average as the ratio of the labeled data increases. Note that the SAD-PL with soft labeling represents a remarkably high average AUC, even with a high proportion of labeled data during training.

Similar to the experiment above, we investigate the effect of including abnormal samples in the unlabeled data during the training with the MNIST, CIFAR10 and MNIST-C datasets. To do this, we increase the anomaly ratio, which is the proportion of abnormal samples in unlabeled training data, from 0% to 10% and represent the average AUCs of 30 times. We also use the one vs. rest evaluation method on the MNIST and CIFAR10 datasets. For the experiment, we keep the labeled data at 95% of the training data and include 98% and 2% of normal and abnormal samples, respectively. Figure 9 shows the variations of performance for the average AUCs of the unsupervised and

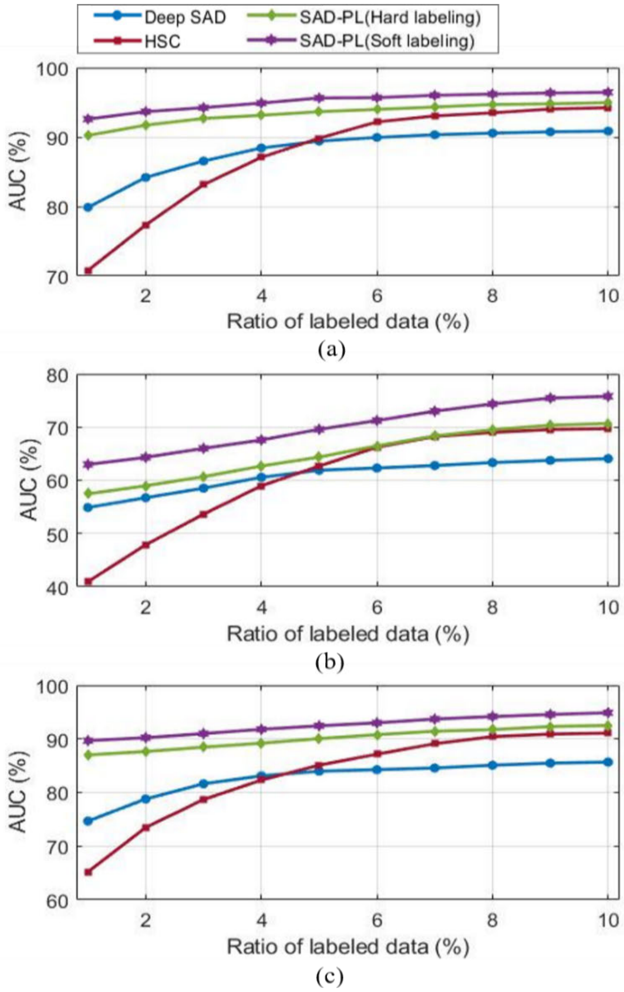
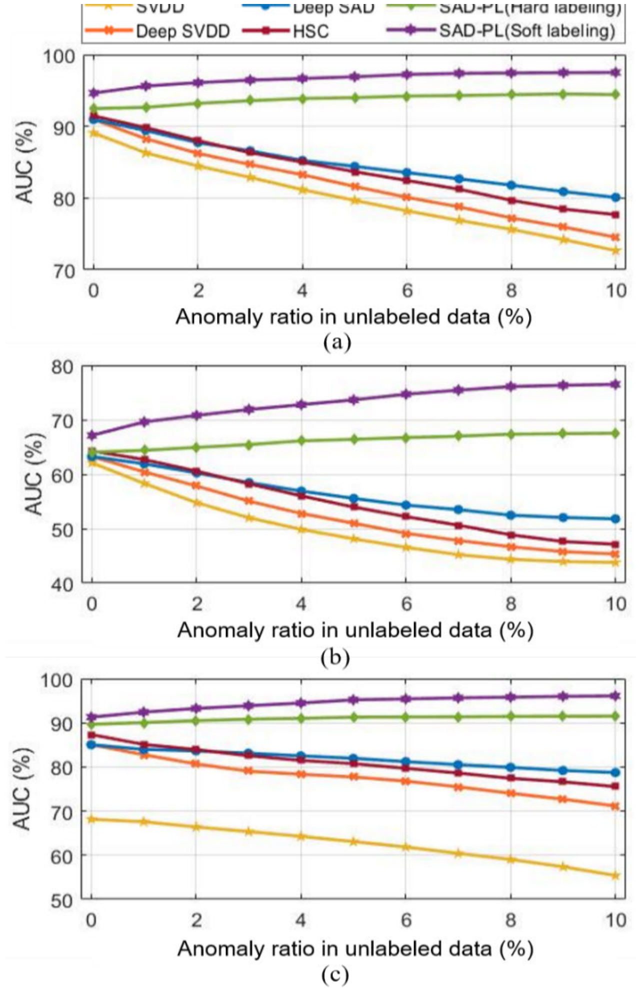


FIGURE 8. Average AUCs according to ratio of labeled data for evaluation on (a) MNIST, (b) CIFAR10, and (c) MNIST-C dataset.



(a)

(b)

(c)

FIGURE 9. Average AUCs according to anomaly ratio in unlabeled data for evaluation using (a) MNIST, (b) CIFAR10, and (c) MNIST-C dataset.

semi-supervised models according to the anomaly ratio in unlabeled data during training. In Figure 9, the unsupervised and semi-supervised approaches assume that the unlabeled data consisting of only normal samples represent performance degradation in the average AUCs as the anomaly ratio in unlabeled data increases. However, the semi-supervised models have a high performance in the average AUC compared to the unsupervised models by learning abnormal samples in labeled data. However, the SAD-PL represents the improvement in performance for the average AUCs as the anomaly ratio in unlabeled data increases via probabilistic labeling. This performance variation appears in both hard and soft labeling.

IV. CONCLUSION

Unlike existing semi-supervised anomaly detection algorithms, which are trained by assuming that most of the samples in unlabeled data are normal, we propose the SAD-PL, which can be applied when abnormal samples are included in unlabeled data. The proposed SAD-PL uses LOF scores obtained from both labeled and unlabeled data and then estimates the labeling probability on the unlabeled data by using the LOF scores. Because of probabilistic labeling and complementary objective function, the SAD-PL has properties of regression and classification. Through experiments, we show that the SAD-PL presents a higher performance in the average AUCs, displays tighter decision boundaries and achieves higher learning efficiency than the existing algorithms. Additionally, the SAD-PL shows an improved performance in the average AUCs as the abnormal data ratio in unlabeled data increases, whereas the existing algorithms show performance degradation. Therefore, the SAD-PL can be a good candidate for providing stable detection performance, regardless of the existence of abnormal samples in unlabeled data. While the proposed SAD-PL shows stable superior detection performance, efficient algorithm for probabilistic labeling without LOF computation is an interesting topic for future research.

REFERENCES

[1] L. Ruff, J. R. Kauffmann, R. A. Vandermeulen, G. Monatavon, W. Samek, M. Kloft, T. G. Dietterich, and K. R. Müller, "A unifying review of deep and shallow anomaly detection," *Proc. IEEE*, vol. 109, no. 5, pp. 756–795, Feb. 2021.

[2] H.-J. Liao, C.-H. R. Lin, Y.-C. Lin, and K.-Y. Tung, "Intrusion detection system: A comprehensive review," *J. Netw. Comput. Appl.*, vol. 36, no. 1, pp. 16–24, Jan. 2013.

[3] M. Ahmed, A. N. Mahmood, and J. Hu, "A survey of network anomaly detection techniques," *J. Netw. Comput. Appl.*, vol. 60, pp. 19–31, Jan. 2016.

[4] Y. Xin, L. Kong, Z. Liu, Y. Chen, Y. Li, H. Zhu, M. Gao, H. Hou, and C. Wang, "Machine learning and deep learning methods for cybersecurity," *IEEE Access*, vol. 6, pp. 35365–35381, 2018.

[5] J. Rabatel, S. Bringay, and P. Poncelet, "Anomaly detection in monitoring sensor data for preventive maintenance," *Expert Syst. Appl.*, vol. 38, no. 6, pp. 7003–7015, Jun. 2011.

[6] L. Martí, N. Sanchez-Pi, J. Molina, and A. Garcia, "Anomaly detection based on sensor data in petroleum industry applications," *Sensors*, vol. 15, no. 2, pp. 2774–2797, Jan. 2015, doi: 10.3390/s150202774.

[7] D. Ramotsoela, A. Abu-Mahfouz, and G. Hancke, "A survey of anomaly detection in industrial wireless sensor networks with critical water system infrastructure as a case study," *Sensors*, vol. 18, no. 8, p. 2491, Aug. 2018, doi: 10.3390/s18082491.

[8] A. Rabaoui, M. Davy, S. Rossignol, and N. Ellouze, "Using one-class SVMs and wavelets for audio surveillance," *IEEE Trans. Inf. Forensics Security*, vol. 3, no. 4, pp. 763–775, Dec. 2008.

[9] H. Lim, J. Park, K. Lee, and Y. Han, "Rare sound detection using 1D convolutional recurrent neural networks," in *Proc. Workshop Detection Classification Acoustic Scenes Events*, 2017, pp. 80–84.

[10] K. Lee and C. H. Lee, "Abnormal signal detection based on parallel autoencoders," *J. Acoust. Soc. Korea*, vol. 40, no. 4, pp. 337–346, Jul. 2021.

[11] H. Yang, K. Lee, Y. Choo, and K. Kim, "Underwater acoustic research trends with machine learning: General background," *J. Ocean Eng. Technol.*, vol. 34, no. 2, pp. 147–154, Apr. 2020.

[12] H. Yang, K. Lee, Y. Choo, and K. Kim, "Underwater acoustic research trends with machine learning: Passive SONAR applications," *J. Ocean Eng. Technol.*, vol. 34, no. 3, pp. 227–236, Jun. 2020.

[13] H. Yang, S.-H. Byun, K. Lee, Y. Choo, and K. Kim, "Underwater acoustic research trends with machine learning: Active SONAR applications," *J. Ocean Eng. Technol.*, vol. 34, no. 4, pp. 277–284, Aug. 2020.

[14] B. Schölkopf, R. Willianson, A. Smola, J. S. Taylor, and J. Platt, "Support vector method for novelty detection," in *Proc. NIPS*, Cambridge, MA, USA, 1999, pp. 582–588.

[15] D. M. J. Tax and R. P. W. Duin, "Support vector data description," *Mach. Learn.*, vol. 54, no. 1, pp. 45–66, Jan. 2004.

[16] L. Ruff, R. A. Vandermeulen, N. Görnits, L. Deecke, S. A. Siddiqui, A. Binder, E. Müller, and M. Kloft, "Deep one-class classification," in *Proc. ICML*, Stockholm, Sweden, 2018, pp. 4393–4402.

[17] Z. Ghafoori and C. Leckie, "Deep multi-sphere support vector data description," in *Proc. SIAM Int. Conf. Data Mining*, Toronto, ON, Canada, 2020, pp. 109–117.

[18] S. Goyal, A. Raghunathan, M. Jain, H. Simhadri, and P. Jain, "DROCC: Deep robust one-class classification," in *Proc. ICML*, 2020, pp. 109–117.

[19] H. Song, Z. Jiang, A. Men, and B. Yang, "A hybrid semi-supervised anomaly detection model for high-dimensional data," *Comput. Intell. Neurosci.*, vol. 2017, Nov. 2017, Art. no. 8501683.

[20] S. Akvay, A. A. Abarghouel, and T. P. Breckon, "GANomaly: Semi-supervised anomaly detection via adversarial training," in *Proc. ACCV*, Perth, WA, Australia, 2018, pp. 622–637.

[21] L. Ruff, R. A. Vandermeulen, N. Görnitz, A. Binder, E. Müller, K. R. Müller, and M. Kloft, "Deep semi-supervised anomaly detection," in *Proc. ICLR*, Addis Ababa, Ethiopia, 2020, pp. 1–23.

[22] D. Hendrycks, M. Mazeika, and T. Dietterich, "Deep anomaly detection with outlier exposure," in *Proc. ICLR*, New Orleans, LA, USA, 2019, pp. 1–18.

[23] L. Ruff, R. A. Vandermeulen, B. J. Franks, K. R. Müller, and M. Kloft, "Rethinking assumptions in deep anomaly detection," in *Proc. ICML*, Jul. 2021, pp. 1–17. [Online]. Available: https://sites.google.com/view/udlworkshop2021/accepted-papers/UDL2021-paper-011.pdf

[24] M. M. Breunig, H. P. Kriegel, R. T. Ng, and J. Sander, "LOF: Identifying density-based local outliers," in *Proc. ACM SIGMOD Int. Conf. Manage. Data*, New York, NY, USA, 2000, pp. 93–104.

[25] J. Neyman and E. S. Pearson, "On the problem of the most efficient tests of statistical hypotheses," *Philos. Trans. Roy. Soc. London A*, vol. 231, nos. 694–706, pp. 289–337, 1933.

[26] J. T. Barron, "A general and adaptive robust loss function," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Long Beach, CA, USA, Jun. 2019, pp. 4331–4339.

[27] Y. Lecun, C. Cortes, and C. Burges. (Jun. 28, 2010). *MNIST Handwritten Digit Database*. AT&T Labs. [Online]. Available: https://yann.lecun.com/exdb/mnist/

[28] A. Krizhevsky. (Apr. 8, 2009). *Learning Multiple Layers of Features From Tiny Images*. [Online]. Available: https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.222.9220&rep=rep1&type=pdf

[29] N. Mu and J. Gilmer, "MNIST-C: Robustness benchmark for computer vision," in *Proc. ICML Workshop*, Jun. 2019, pp. 1–11. [Online]. Available: https://www.gatsby.ucl.ac.uk/ balaji/udl2019/accepted-papers/UDL2019-paper-37.pdf

[30] A. F. Emmott, S. Das, T. Dietterich, A. Fern, and W.-K. Wong, "Systematic construction of anomaly detection benchmarks from real data," in *Proc. ACM SIGKDD Workshop Outlier Detection Description (ODD)*, Chicago, IL, USA, 2013, pp. 16–21.

KIBAE LEE (Graduate Student Member, IEEE) received the B.S. degree and M.S. degree in ocean system engineering from Jeju National University, in 2016 and 2017, respectively, where he is currently pursuing the Ph.D. degree with the Department of Ocean System Engineering. His research interests include machine learning, acoustic signal processing, and sonar systems.

CHONG HYUN LEE (Member, IEEE) received the B.S. degree in electronic engineering from Hanyang University, in 1985, the M.S. degree from Michigan Technological University, in 1987, and the Ph.D. degree in electronic engineering from Korea Advanced Institute of Science and Technology (KAIST), in 2002. He is currently working as a Professor with the Department of Ocean System Engineering, Jeju National University. His research interests include machine learning, sonar signal processing, and smart RF sensor design.

JONGKIL LEE received the B.S. degree in mechanical engineering from Pusan National University, in 1984, and the M.S. degree and the Ph.D. degree in mechanical engineering from The University of Utah, in 1990 and 1993, respectively. He is currently working as a Professor with the Department of Mechanical Engineering Education, Andong National University. His research interests include underwater acoustics, noise and vibration, fiber optic sensors, dynamics, and energy harvesting systems.